R4 サウンドとジェスチャの学習に基づくダンスの生成

下山功　山本正信
新潟大学大学院自然科学研究科

1 はじめに

音楽番組で曲に合わせてダンスを踊る影像はよく見掛けるが、あらかじめ練習した振り付けに従って踊るのではなくその時の曲の雰囲気などを掴んで独特な動きで踊る歌手もよく見掛ける。また、テレビやラジオで聞くようなものでなくても歌に合わせて独特な踊りをするものは日本の能や歌舞伎など多数ある。

これらの動きをCGで再現するには、あらかじめタイミングがマッチするようにジェスチャとサウンドのデータを与えておく必要があるがマッチングには多大な手間と時間を必要とする。

そこで歌手のジェスチャとサウンドの関連付けを計算機に学習させサウンドを与えるだけでその歌手の動きを再現できるようにしておけばある歌手のジェスチャを他の曲、さらにはジャンルの異なる曲に対しても少ない手間と時間で生成することができる。また現在の音楽に対して過去の歌手がどのような動きをするか計算機上で実現することも可能となる。

実現方法としてはVideo rewrite [1] やVoice puppetry [2] 等があるが前者は画像中のキャラクタではなく画像そのものを生成するものであり、後者は顔の表情の生成はできるが体全体のジェスチャの生成には至っていない。

本研究ではある時点での動作は過去のジェスチャ、過去のサウンドの経験と未来のサウンドの予測から生成されるという仮定を基にサウンドとジェスチャの関連付けの学習に自己回帰モデルを用いる方法を提案する。まず、画像中のキャラクタのジェスチャをトラッキングして関節角度の時系列データ列として測定する。さらにそれに対応する音声データ列も取り出し、この2つのデータ列の関連付けを行う。得られた回帰係数に対し新しい音声データ列を与えるとその新しい音に対する動作を生成する。

2 サウンドとジェスチャの関係の学習と動作の生成

対応付けのモデルは y_n を動作データ列、x_{ni} を音声データ列、a_i を回帰係数、b_j を自己回帰係数とするとき

$$y_n = \sum_{i=1}^{m} a_i x_{ni} + \sum_{j=1}^{l} b_j y_{n-j} \qquad (1)$$

となる。この式を用いて関連付けの学習を行う(自己回帰関係を考慮しないものとして[5]がある)。まず、動作データは図1のような多関節モデルを用いて画像中のキャラクタの動作追跡を行い関節角度の時系列データとして測定する（動作測定の詳細は文献[3][4]を参照されたい)。

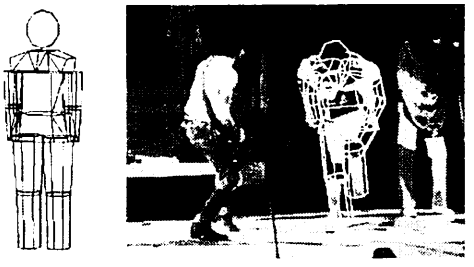

図1: (左) 人体幾何モデル、(右) 動作の測定.「THE ROLLING STONES LET'S SPEND NIGHT TOGETHER」©1983 Promotour B.V. All Rights Reserved.

また音声データは、44100Hzのレートでサンプリングを行った。例えば1秒間に30フレームの画像があるとするならば、1フレーム間には1470点の音声データが含まれていることになる。本研究では1フレームでのサンプリングデータ2乗和の平均、すなわち音のパワーを1フレームでの音声データとして与えることとする。関節角度を出力データ、音のパワーを入力データとしてモデルに与えると係数 a_i, b_j を最

小二乗法より決定することができる。求めた回帰係数を用いて新たな音声データ列を式(1)に与えることで動作データ列の生成ができる。

以下に関係の学習と動作生成の例を挙げる(自己回帰関係を考慮しない結果を示す)。図2の(a),(b)のデータ列より回帰係数を求める。求めた係数を用いて式(1)に図2の(c)の新しい音声データ列を与える。その結果が図2の(d)の動作データ列である。

図2の(a)(b)の間には(c)(d)ほど強い相関は見られない。つまり、歌手の動作には音のパワー意外の様々な要因が考慮されているものと考えられる。動作生成の要因としては音の周波数やリズムなどいろいろなものが考えられるが本研究では動作生成のための新たな要因として自己回帰関係を導入した。

3 むすび

ジェスチャを生成するための要因としてある時点とその前後の区間のサウンドに加え自己回帰関係を考慮にいれた関係の学習を行った。そして得られた関係に新たな音声データを与えて動作生成を行った。なお、この報告の時点では左の太股を上げる等の単一の動作の学習だけを行った。なぜなら、複数の動作を学習させるとその複数の動作を1度に生成させるような学習を行ってしまうためそこから得られた関係からジェスチャを生成しても平均化された動き、極端になるとほとんど動かないようになってしまう。そこで、今後の課題は並進移動を含めた複数の動作の生成が挙げられる。また、その他の動作生成の要因、たとえばサウンドの周波数やリズムなどを併用することも考えられる。

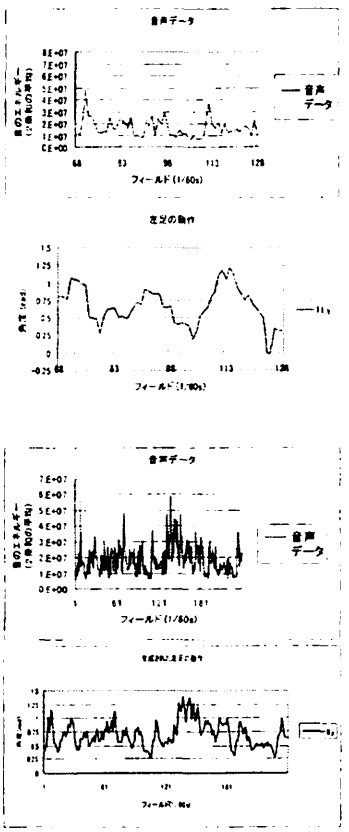

図2: 上から順に(a)元の音声データ、(b)元の動作データ、(c)新しく与えた音声データ、(d)新しく生成された動作データ、動作データは(b)(c)共に測定した人物の左脚の関節角度のデータ列

参考文献

[1] C.Bregler,M.CovellandM.Slaney,"Video rewrite: Driving visual speech with audio",in Proc.ACM SIGGRAPH'97,pp353-360,1997

[2] M.Brand,"Voicepuppetry",inProc.ACM SIGGRAPH'99,pp21-28,1999

[3] 大田佳人、山際隆志、山本正信：キーフレーム拘束を利用した単眼動画像からの人間動作の追跡、電子情報通信学会論文誌、Vol.J81-D-II、No.9,pp.2008-2018,1998

[4] 山本正信、川田聡、近藤拓也、越川和忠：ロボットモデルに基づく人間動作の3次元動画像追跡、電子情報通信学会論文誌、Vol.J79-D-II, No.1,pp.71-83,1996

[5] 山本正信、星昌人、下山功、五十嵐達也：サウンドとジェスチャの対応付けからのキャラクタの動作生成、電子情報通信学会技術研究報告、PRMU 2001-8～19、pp.27-32,2001